০০১-০০২ Statistics এর ভূমিকা (পর্ব-১)

Statistics কী?

Statistics (পরিসংখ্যান) হলো একটি গণিতশাস্ত্রের শাখা যা ডেটা সংগ্রহ, বিশ্লেষণ, উপস্থাপন, এবং ব্যাখ্যার মাধ্যমে অর্থপূর্ণ সিদ্ধান্ত গ্রহণে সহায়তা করে।

Statistics মূলত কত প্রকার?

দুই প্রকার। যথাঃ

১) Descriptive Statistics

২) Inferential Statistics

[ছবিটি[1] পাব্লিশ করেছেন **alejo15084**. আপনি তাকে ডোনেট করতে পারেন।]

---

[1] https://pixabay.com/illustrations/ai-generated-dashboard-statistics-8615473/

Descriptive Statistics কী?

Descriptive Statistics এমন একটা শাখা যা ডেটাকে সারসংক্ষেপ করে এবং এমনভাবে উপস্থাপন করে যাতে তা সহজে বোঝা যায়। এটি কেবলমাত্র সংগৃহীত ডেটার বিশ্লেষণ করে এবং নতুন কোনো সিদ্ধান্ত অনুমান করে না।

উদাহরণ:

১) একটি স্কুলের ১০০ জন ছাত্রের পরীক্ষার নম্বর বিশ্লেষণ করা।

২) গড় (Mean), মধ্যক (Median), প্রচুরক (Mode) বের করা।

৩) বার চার্ট, পাই চার্ট, হিস্টোগ্রাম, এবং টেবিলের মাধ্যমে ডেটা উপস্থাপন করা।

মূলতঃ Descriptive Statistics কেবল ডেটার সংক্ষিপ্ত চিত্র উপস্থাপন করে, কিন্তু ভবিষ্যদ্বাণী করে না।

Inferential Statistics কী?

Inferential Statistics এমন একটি শাখা যা নমুনা ডেটার ওপর ভিত্তি করে সম্পূর্ণ জনসংখ্যা সম্পর্কে সিদ্ধান্ত গ্রহণ করে। এটি ভবিষ্যৎ অনুমান বা পূর্বাভাস দিতে ব্যবহৃত হয়।

উদাহরণ:

১) ১০০ জন ছাত্রের নমুনার ওপর ভিত্তি করে সম্পূর্ণ দেশের ছাত্রদের গড় পরীক্ষার নম্বর অনুমান করা।

২) কোনো ওষুধের কার্যকারিতা পরীক্ষার জন্য ৫০০ জন রোগীর ওপর গবেষণা করে সেটির প্রভাব সম্পর্কে সিদ্ধান্ত নেওয়া।

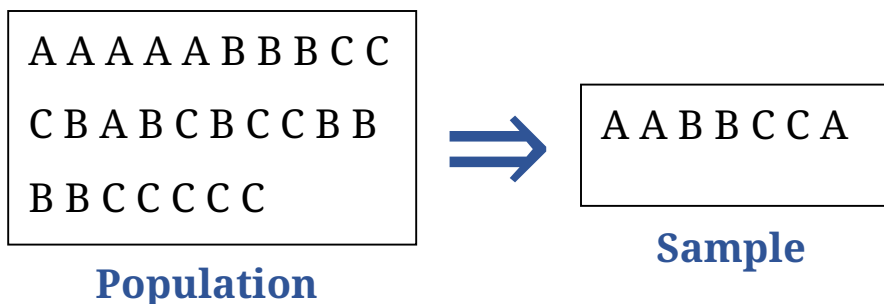
Inferential Statistics -এর গুরুত্বপূর্ণ উপাদান:

1. হাইপোথিসিস টেস্টিং (Hypothesis Testing):
   a. কোনো অনুমানের সত্যতা যাচাই করা (যেমন, "নতুন ওষুধটি কার্যকর কি না")।
   b. t-Test, Chi-Square Test, ANOVA ইত্যাদি ব্যবহার করা হয়।
2. বিশ্বাসযোগ্যতা এবং প্রভিশনের অনুমান (Confidence Intervals & Predictions):
   a. একটি নির্দিষ্ট পরিসীমার মধ্যে ভবিষ্যদ্বাণী করা। উদাহরণ: "৭০%-৯০% সম্ভাবনা রয়েছে যে ২০২৫ সালে গড় তাপমাত্রা ২°C বৃদ্ধি পাবে।"
3. রিগ্রেশন অ্যানালাইসিস (Regression Analysis):
   a. দুই বা ততোধিক ভেরিয়েবলের মধ্যে সম্পর্ক বিশ্লেষণ করা। উদাহরণ: "বিজ্ঞাপনের খরচ বাড়ালে বিক্রি কতটা বাড়তে পারে?"

Population কী?

Population হলো সমগ্র data এর কালেকশন। যেমন: "কখগ" স্কুলে ১ম শ্রেণী থেকে ১০ম শ্রেণী পর্যন্ত সকল শিক্ষার্থীদের ডেটা।

Sample কী?

Sample হলো Population এর Subgroup। যেমন: "কখগ" স্কুলে শুধুমাত্র ১০ম শ্রেণীর শিক্ষার্থীদের ডেটা। অথবা, "কখগ" স্কুলে ১ম শ্রেণী থেকে ১০ম শ্রেণী পর্যন্ত সকল শিক্ষার্থীদের থেকে মাত্র ১০০ জনের ডেটা।

Descriptive Statistics আর Inferential Statistics এর উপর উদাহরণ:

ধরুন, "সানশাইন লাইটিং" নামের একটি কোম্পানি প্রতি মাসে ৫০,০০০টি LED বাল্ব উৎপাদন করে। কোম্পানিটি বাল্বের গুণগত মান নিশ্চিত করতে তাদের উৎপাদিত বাল্বগুলোর আলোর উজ্জ্বলতা (Brightness) পরিমাপ করতে চায়, যা লুমেন (Lumen) এককে প্রকাশ করা হয়।

কোম্পানি যদি পুরো ৫০,০০০টি বাল্বের উজ্জ্বলতা পরিমাপ করে এবং নিচের ফলাফলগুলো সংগ্রহ করে, তাহলে এটি বর্ণনামূলক পরিসংখ্যান হবে—

সর্বোচ্চ উজ্জ্বলতা: ৯৮০ লুমেন

সর্বনিম্ন উজ্জ্বলতা: ৭৮০ লুমেন

গড় উজ্জ্বলতা: ৮৫০ লুমেন

মধ্যক (Median): ৮৪৫ লুমেন

বিচ্যুতি (Standard Deviation): ৫০ লুমেন

এখানে আমরা শুধুমাত্র সংগৃহীত তথ্যের সারাংশ তৈরি করেছি। কোনো ভবিষ্যদ্বাণী করিনি বা অনুমান করিনি। অর্থাৎ এটি হলো Descriptive Statistics

কিন্তু ৫০,০০০টি বাল্বের প্রতিটি পরীক্ষা করা ব্যয়বহুল এবং সময়সাপেক্ষ। তাই কোম্পানি এলোমেলোভাবে ৫০০টি বাল্বের নমুনা নিয়ে পরীক্ষা চালায় এবং নিচের ফলাফল পায়—

গড় উজ্জ্বলতা: ৮৫২ লুমেন

বিচ্যুতি: ৪৮ লুমেন

এখন, কোম্পানি এই নমুনা ডাটার ভিত্তিতে অনুমান করে যে—

"আমাদের সমস্ত ৫০,০০০টি বাল্বের গড় উজ্জ্বলতা আনুমানিক ৮৫০ ± ৫ লুমেনের মধ্যে থাকবে।"

"প্রায় ৯৫% বাল্বের উজ্জ্বলতা ৮৫০ লুমেনের আশেপাশে থাকবে।"

এখানে, একটি ছোট নমুনার (Sample) ভিত্তিতে সমস্ত উৎপাদিত বাল্বের (Population) গুণগত মান সম্পর্কে অনুমান করা হয়েছে। তাই এটি Inferential Statistics।

Descriptive Statistics বনাম Inferential Statistics পার্থক্য:

বৈশিষ্ট্য	Descriptive Statistics	Inferential Statistics
উদ্দেশ্য	সংগৃহীত ডেটা বিশ্লেষণ ও উপস্থাপন করা	নমুনা ডেটার ভিত্তিতে ভবিষ্যদ্বাণী করা
উপাদান	গড়, মধ্যক, প্রচুরক, পরিসর, মানক বিচ্যুতি	হাইপোথিসিস টেস্টিং, রিগ্রেশন অ্যানালাইসিস, ভবিষ্যৎ পূর্বাভাস
সিদ্ধান্ত গ্রহণ	শুধুমাত্র সংগৃহীত ডেটা নিয়ে কাজ করে	জনসংখ্যা সম্পর্কে অনুমান করে
ব্যবহার	পরীক্ষার ফলাফল বিশ্লেষণ, গ্রাফ ও চার্ট তৈরি	ওষুধের কার্যকারিতা নির্ণয়, জনমত জরিপের পূর্বাভাস

ডেটা কি?

ডেটা (Data) হল কোনো নির্দিষ্ট প্রসঙ্গে সংগ্রহ করা কাঁচা তথ্য, সংখ্যা, অক্ষর, চিত্র বা যেকোনো উপাত্ত যা বিশ্লেষণ করা সম্ভব।

যেমন:

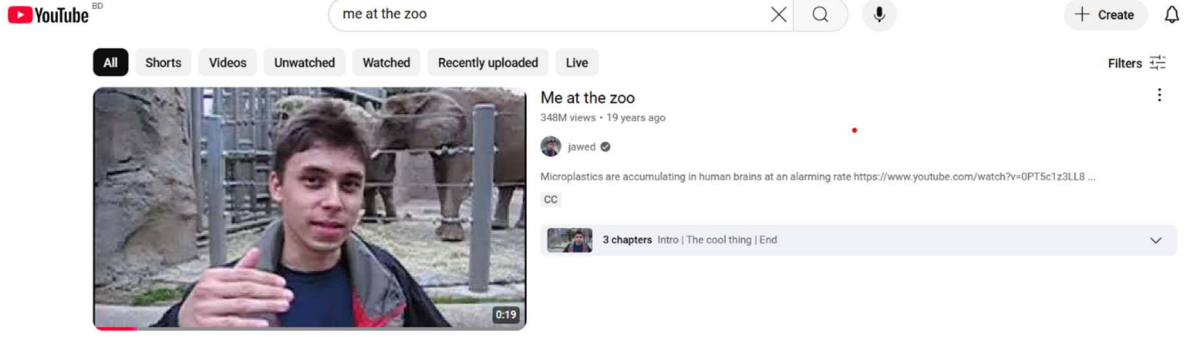
24 January 2025

Marketing cost : 480 Tk

Travel cost : 60 Tk

এটা হল ডেটা।

Statistics নির্ভর করে ডেটা এর উপর।

বর্তমানে ডেটা মানে সবকিছুই বুঝানো হয়, যেমনঃ ইউটিউবের ভিডিও, ভিডিওর কমেন্ট, লাইক, ইন্টারনেটের যেকোনো জিনিস, উইকিপেডিয়ার আর্টিকেল ইত্যাদি।

চিত্রঃ ইউটিউবের co-founder জাওয়েদ করিমের প্রথম ভিডিও "Me at the zoo"। এই ভিডিওকেও ডেটা বলা হয়।

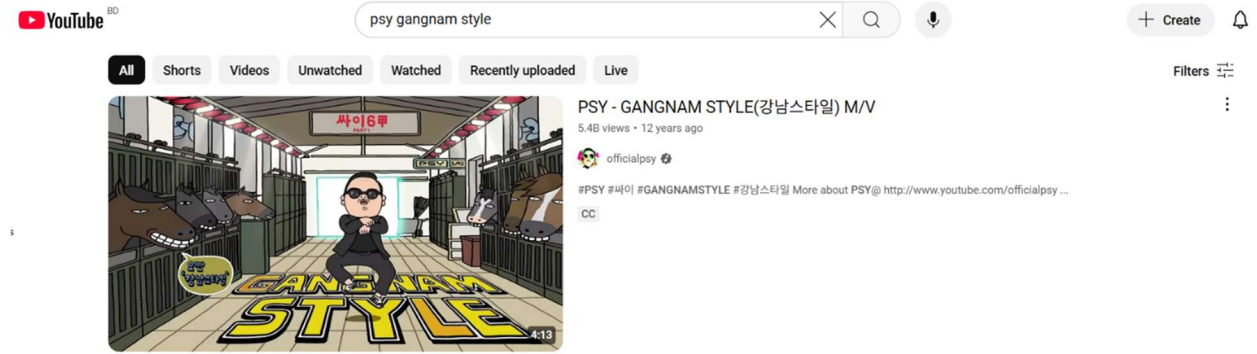
Reference Link: https://youtu.be/jNQXAC9IVRw?si=Tpc8cb1gCf52dOPs

চিত্রঃ উইকিপেডিয়াতে আলবার্ট আইনস্টাইনের আর্টিকেল। এটাও হল ডেটা। Reference Link: https://en.wikipedia.org/wiki/Albert_Einstein

ডেটার ওপর বিভিন্ন বিশ্লেষণ ও প্রক্রিয়াকরণের মাধ্যমে যখন তা অর্থবহ হয়, তখন তাকে ইনফরমেশন বলা হয়। এবং ঐ ইনফরমেশন নিয়ে আমরা কাজ করে থাকি।

আমাদের কেন ডেটা সংগ্রহ করতে হয়?

১) কোনো নির্দিষ্ট দল বা ব্যক্তি, স্থান, বা বস্তুর বৈশিষ্ট্য বিশ্লেষণ করার জন্য আমাদের ডেটা সংগ্রহ করতে হয় (যেমনঃ "কখগ" স্কুলে ১ম শ্রেণী থেকে ১০ম শ্রেণী পর্যন্ত সকল শিক্ষার্থীদের থেকে মাত্র ১০০ জনের পরিক্ষার রেজাল্টের ডেটা বিশ্লেষণ করে স্কুলের চারিত্রিক বৈশিষ্ট্য গবেষণা করা)

২) কোন ইউটিউব মিউজিক ভিডিও সবচেয়ে বেশি পপুলার হইছে, সেইটাও ডেটার মাধ্যমে জানা যায়।

চিত্রঃ PSY এর Gangnam Style। এই গান এত জনপ্রিয় ছিল যে ৫.৪ বিলিয়ন ভিউ হয়। Reference Link: https://youtu.be/9bZkp7q19f0?si=XYD-BYPT_3qyDFbD

ডেটা সংগ্রহঃ

ডেটা available থাকলেঃ ডেটা প্রকাশ করা হবে।

ডেটা available না থাকলেঃ ডেটা সংগ্রহ বা তৈরি করতে হবে।

স্ট্রাকচারের ভিত্তিতে ডেটার প্রকারভেদ:

স্ট্রাকচারের ভিত্তিতে ডেটা দুই প্রকারঃ

১) Unstructured data

2) Structured data

Unstructured data কী? আর Structured data কী?

যেসব ডেটা নির্দিষ্ট বিন্যাস বা কাঠামো তে নাই, তারা হলো Unstructured data। আর যেসব ডেটা নির্দিষ্ট বিন্যাস বা কাঠামো তে আছে, তারা হলো Strucutred data।

উদাহরণঃ

একটা দোকানে কাস্টমাররা কি কি দ্রব্য খরিদ করল, তার ডেটাঃ

কাস্টমার ১: ল্যাপটপ, ইউএসবি ড্রাইভ, ওয়্যারলেস মাউস ।

কাস্টমার ২: স্মার্টওয়াচ, এইচডিএমআই কেবল।

কাস্টমার ৩: ব্লুটুথ স্পিকার, গেমিং কীবোর্ড, এসএসডি

এই ডেটা হলো Unstructured data, কারণ এই ডেটা নির্দিষ্ট বিন্যাস বা কাঠামো তে নাই। এই ডেটা থেকে ইনফরমেশন বের করা তুলনামূলক কঠিন। (যেমনঃ “কাস্টমার ১” তো ল্যাপটপ, ইউএসবি ড্রাইভ, ওয়্যারলেস মাউস খরিদ করলো, তাতে ঐ কাস্টমার কত টাকা দোকানদার কে দিল, সেই ইনফরমেশন পাওয়া যাচ্ছে না।)

কিন্তু কাস্টমাররা কি কি দ্রব্য খরিদ করল, সেই ডেটা যদি এরকম হতোঃ

কাস্টমার ১	
পণ্য	মুল্য (টাকা)
ল্যাপটপ	৬০০০০
ইউএসবি ড্রাইভ	৬০০
ওয়্যারলেস মাউস	২৯৫০

কাস্টমার ২	
পণ্য	মুল্য (টাকা)
স্মার্টওয়াচ	১৪৯০
এইচডিএমআই কেবল	৪৮০

কাস্টমার ৩	
পণ্য	মুল্য (টাকা)
ব্লুটুথ স্পিকার	২৩৫০
গেমিং কীবোর্ড	৭৫০
এসএসডি	১৪০০

তখন এই ডেটা হবে Structured data, কারণ এই ডেটা নির্দিষ্ট বিন্যাস বা কাঠামোতে আছে। এখান থেকে ইনফরমেশন বের করা যায়, যেমনঃ “কাস্টমার ১” যেসব দ্রব্য কিনলো, সেক্ষেত্রে তার কাছে দোকানদার পায় (৬০০০০+৬০০+২৯৫০) টাকা = ৬৩৫৫০ টাকা।

Structured data এর আরেকটা উদাহরণ, এখানে ১০ জন শিক্ষার্থীর নাম, রোল আর মার্কস দেওয়া আছে।

	A	B	C
1	Name	Roll Number	Marks
2	Olivia C.	1	99
3	Alice J.	2	93
4	Frank Z.	3	91
5	Tom F.	4	91
6	Hannah X.	5	91
7	Ryan A.	6	80
8	Grace U.	7	85
9	Mia U.	8	75
10	Sophia Q.	9	83

এই ডেটা হলো Structured data কারণ এটা নির্দিষ্ট বিন্যাস বা কাঠামোতে এ আছে। এখানে আমরা সহজে ইনফরমেশন পাচ্ছি। Structured data কে আরেকভাবে বলা যায় Tabulated data।

ইউটিউব, ফেসবুক, ইন্সটাগ্রাম – ইত্যাদির কমেন্টস, পোস্ট হল Unstructured data। যেমনঃ কেউ ইউটিউবে কমেন্ট করলো।

“The song is beautiful ever. The artist is also nice.”

এটা হল Unstructured data। একে structured data তে আনতে হলে এভাবে আনতে হবে।

Person/Object	**State**
Song	Beautiful
Artist	Nice

ঠিক এভাবে আমরা Unstructured data কে Structured data তে কনভার্ট করতে পারি।

ডেটাবেসে সংরক্ষিত তথ্য (ইনফরমেশন) তখনই কার্যকর হয় যখন আমরা সেই তথ্যের প্রাসঙ্গিকতা ও অর্থ বুঝতে পারি। যখন সংখ্যা বা টেক্সট এলোমেলোভাবে ছড়িয়ে থাকে এবং কোনো নির্দিষ্ট কাঠামো বা সংগঠন থাকে না, তখন সেই তথ্য থেকে কার্যকর সিদ্ধান্ত নেওয়া বা বিশ্লেষণ করা কঠিন হয়ে পড়ে। তাই, ডেটাবেসে তথ্যকে সুসংগঠিত ও কাঠামোবদ্ধভাবে সংরক্ষণ করা অত্যন্ত গুরুত্বপূর্ণ।

Dataset:

Dataset হলো Structured collection of data। যেমন:

Customer Name	Buying Amount	Total price (Tk)
Arif Rahman	2	200
Mojid Mollah	10	2660
Akash	6	1450
Karim Uddin	5	1330

এটা হলো ডেটাসেট। আর এটাই হলো Structured data.

Variables and Cases:

ভেরিয়েবল (Variable) কী?

স্ট্যাটিস্টিক্সে ভেরিয়েবল (Variable) হলো একটি বৈশিষ্ট্য বা গুণ যা পরিবর্তনশীল হতে পারে। এটি বিভিন্ন ব্যক্তি, বস্তু বা ঘটনাসমূহের মধ্যে পৃথক মান ধারণ করতে পারে।

উদাহরণ:

- একজন ছাত্রের বয়স, উচ্চতা, ওজন (যেগুলো ব্যক্তি ভেদে পরিবর্তিত হতে পারে)।
- কোনো পণ্যের দাম বা বিক্রির পরিমাণ (যেগুলো সময়ের সাথে পরিবর্তিত হতে পারে)।

ভেরিয়েবল দুই প্রকার:

- গুণগত বা ক্যাটাগরিকাল ভেরিয়েবল (Qualitative or Categorical Variable) – যেমন লিঙ্গ (পুরুষ/নারী), রক্তের গ্রুপ (A, B, O)।

- পরিমাণগত বা নিউমেরিক্যাল ভেরিয়েবল (Quantitative or Numerical Variable) – যেমন বয়স (২০ বছর), উচ্চতা (৫.৬ ফুট)।

কেস (Case) কী?

স্ট্যাটিস্টিক্সে কেস (Case) বলতে এমন একটি একক বা অবজেক্টকে বোঝানো হয়, যার জন্য আমরা তথ্য সংগ্রহ করি। প্রতিটি কেসের এক বা একাধিক ভেরিয়েবল থাকতে পারে।

উদাহরণ:

- একটি স্কুলের ছাত্রদের নিয়ে গবেষণা করা হলে, প্রতিটি ছাত্র একটি কেস হবে। ডেটাসেটে ছাত্রদের নাম হবে Case।
- একটি কোম্পানির বিভিন্ন পণ্যের বিক্রয় বিশ্লেষণ করলে, প্রতিটি পণ্য একটি কেস হবে। ডেটাসেটে পণ্যের নাম হবে Case

কেস সাধারণত সারি (Row) আকারে উপস্থাপন করা হয়, যেখানে প্রতিটি কেসের জন্য বিভিন্ন ভেরিয়েবলের মান লিপিবদ্ধ থাকে।

উদাহরণ:

ব্যক্তি (কেস)	বয়স (ভেরিয়েবল)	উচ্চতা (ভেরিয়েবল)	লিঙ্গ (ভেরিয়েবল)
রাহুল	২০ বছর	৫.৭ ফুট	পুরুষ
সোহা	২২ বছর	৫.৩ ফুট	নারী
হাসান	২৫ বছর	৫.৯ ফুট	পুরুষ

এখানে প্রতিটি সারি একটি "কেস", এবং প্রতিটি কলাম একটি "ভেরিয়েবল"।

এখানে “রাহুল” হোলো কেস, কারণ আমরা রাহুল কে কেন্দ্র করে তার বয়স, উচ্চতা আর লিঙ্গ বের করছি। একই জিনিসটা সোহা, হাসানের ক্ষেত্রেও যায়। কাজেই “সোহা”, “হাসান” হলো কেস (Case)

Constant (ধ্রুবক):

Constant হলো যা অপরিবর্তনীয়। যেমন:-

	A	B	C	D
1	Name	Roll Number	Marks	Fee Payment
2	Olivia C.	1	99	2000
3	Alice J.	2	93	2000
4	Frank Z.	3	91	2000
5	Tom F.	4	91	2000
6	Hannah X.	5	91	2000
7	Ryan A.	6	80	2000
8	Grace U.	7	85	2000
9	Mia U.	8	75	2000
10	Sophia Q.	9	83	2000

এখানে Fee Payment হলো constant, কারণ সবার Fee Payment একই।

শূণ্য বনাম খালি:

মনে করেন, ১০ মার্কের Quiz পরীক্ষার রেজাল্ট দেওয়া হয়েছে। দেখা গেছে, সেখানে আফ্রিদি পেয়েছে ৯, লাবিবা পেয়েছে ৭, রাকিব পেয়েছে ০, আর মুহিতের মার্ক আসেনি।

নাম	মার্কস (১০)
আফ্রিদি	৯
রাকিব	০
মুহিত	---
লাবিবা	৭

মুহিতের মার্ক খালি মানে মুহিত শূণ্য পায়নি। শূণ্য আর খালি এক জিনিস নয়। মুহিত পরীক্ষা দেয়নি, তাই মুহিতের মার্কস খালি। কিন্তু রাকিব তো পরীক্ষা দিয়েছে, সেই পরীক্ষায় শূণ্য পেয়েছে।

অনেক সময় খালি ডেটা কে “Not available”, “N/A” দ্বারা প্রকাশ করা হয়। খালি ডেটা কে Not available ডেটা বলে। যে ডেটা আমরা পাইনি, সেটাকে Not available ডেটা বা খালি ডেটা বলে।

আমরা স্ট্যাটিস্টিকস কে কিছুক্ষণ বিরতি রেখে পাইথন প্রোগ্রামিং শিখবো। এরপর আবার স্ট্যাটিস্টিকসে ফিরে আসবো।